As per the Latest Syllabus of Anna University, Chennai (Regulation-2017)

BIG DATA ANALYTICS

For B.E./B.Tech VI Semester IT &
VII Sem CSE Branches (Professional Elective)

ARS PUBLICATIONS
CHENNAI.

S. ARUNPRASATH
K. SRIRAM KUMAR
P. KRISHNA SANKAR

PREFACE

This book “Big Data Analytics” is to know about the fundamental concepts of big data, streams and analytics, with various tools and practices in real world. It contributes an impression towards big data programming concepts of R and Python. It provides a preliminary study to access and perform analytics on huge volume of data. It affords procedural footsteps and study over NoSQL, Twitter data analytics and Wikipedia blog.

Unit I: Introduction towards evolution, best practices and characteristics of Big Data. Outline about use cases on Bigdata storage and architecture, real world Hadoop analytics mechanism available currently.

Unit II: Outline towards clustering, K-means and procedural steps in cluster construction. Classification and its core mechanism decision tree, Naïve Bayes are systematically briefed with R program.

Unit III: Transient awareness on association rules, Apriori algorithm and recommendation system. Brief knowledge over detecting candidate rules and collaborative, Content based, Knowledge based and Hybrid Recommendation with applications.

Unit IV: Contributes a knowledge on trendy Stream computing and its architecture. Real world analytics like Sentiment analysis on Twitter, Stock market prediction and Graph analytics are briefed with understanding them over stream computing.

Unit V: Provides a study over NoSQL and various real-world methodology. Various case studies on Hive and Hadoop architecture used in Twitter, E-Commerce and Blogs are briefed. It provides introduction towards R programming and its available function to perform data analytics.

Contents

UNIT I

INTRODUCTION TO BIG DATA

UNIT II

CLUSTERING AND CLASSIFICATION

2.1.2. K-means

2.1.2.1. Use Cases

2.1.2.2. Overview of the Method

2.1.2.3. Determining the Number of Clusters

2.1.2.4. Diagnostics

2.1.2.5. Reasons to Choose and Cautions

2.2. Classification

2.2.1. Decision Trees

2.2.1.1. Overview of a Decision Tree

2.2.1.2. General Algorithm

2.2.1.3. Decision Tree Algorithms

2.2.1.4. Evaluating a Decision Tree

2.2.1.5. Decision Trees in R

2.2.2. Naïve Bayes

2.2.2.1. Baye’s Theorem

2.2.2.2. Naïve Bayes Classifier

UNIT III

ASSOCIATION AND RECOMMENDATION SYSTEM

3.1. Advanced Analytical Theory and Methods: Association Rules

3.1.1. Overview

3.1.2. Apriori Algorithm

3.1.3. Evaluation of Candidate Rules

3.1.4. Applications of Association Rules

3.1.5. Finding Association and Finding Similarity

3.2. Recommendation System

3.2.1. Collaborative Recommendation

3.2.2. Content Based Recommendation

3.2.3. Knowledge Based Recommendation

3.2.4. Hybrid Recommendation Approaches

3.2.5. Demographic Filtering

3.2.6. Evaluation of Recommender Systems

Unit IV

Stream Memory

4.1 Introduction to Streams Concepts

4.2 Stream Data Model and Architecture

4.2.1 A Data-Stream-Management System

4.2.2 Examples of Stream Sources

4.2.3 Stream Queries

4.3 Stream Computing

4.3.1 Data Stream Processing Platforms

4.3.2 Stream computing use cases

4.3.3 Few cross-industry scenarios best suitable for stream computing

4.4 Sampling Data in a Stream

4.4.1 Motivating Example

4.4.2 Obtaining a Representative Sample

4.4.3 General Sampling Problem

4.4.4 Varying the Sample Size

4.5 Filtering Streams

4.5.1 Motivating Example

4.5.2 Bloom Filter

4.5.3 Analysis of Bloom Filtering

4.6 Counting Distinct Elements in a Stream

4.6.1 Count-Distinct Problem

4.6.2 Flajolet-Martin Algorithm

4.6.3 Combining Estimates

4.7 Estimating moments

4.7.1 Definition of Moments

4.7.2 Alon-Matias-Szegedy Algorithm for Second Moments

4.7.3 Why the Alon-Matias-Szegedy Algorithm Works

4.7.4 Higher-Order Moments

4.7.5 Dealing With Infinite Streams

4.8 Counting oneness in a Window

4.8.1 Cost of Exact Counts

4.8.2 Datar-Gionis-Indyk-Motwani Algorithm

4.8.3 Storage Requirements for the DGIM Algorithm

4.8.4 Query Answering in the DGIM Algorithm

4.8.5 Maintaining the DGIM Conditions

4.8.6 Reducing the Error

4.8.7 Extensions to the Counting of Ones

4.9 Decaying Window

UNIT V

NOSQL DATA MANAGEMENT FOR BIG DATA AND VISUALIZATION

5.5.2. HBase Architecture

5.5.3. Master Server

5.5.4. Regions

5.5.5. Region server

5.5.6. Zookeeper

5.6. Analyzing - Big data with twitter

5.6.1. Who is Influential?

5.6.2. About Twitter

5.6.3. Steps to index the Twitter Data in Splunk

5.6.4. Searching Twitter data

5.7. Analyzing - Big data for E-Commerce

5.7.1. Recommendation Engine

5.7.2. Personalized Shopping Experience

5.7.3. Everything in the cart is tracked (SimonVera/Shutterstock)

5.7.4. Voice of the Customer

5.7.5. Dynamic Pricing

5.7.6. Demand Forecasting

5.8. Analyzing - Big data for blogs

5.8.1. End user - working mechanism

5.8.2. Software Architecture

5.8.3. Wikistats – Open Source UI – Community enhancement

5.8.4. MediaWiki

5.9. Review of Basic Data Analytic Methods using R

5.9.1. Introduction to R

5.9.2. Exploratory Data Analysis

5.9.3. Statistical Methods for Evaluation